Worst-Case Execution Time Analysis

Prof. Dr. Jian-Jia Chen

LS 12, TU Dortmund

04, Dec., 2019
Most Essential Assumptions for Real-Time Systems

**Upper bound on the execution times:**

- Commonly, called the *Worst-Case Execution Time (WCET)*
What does Execution Time Depend on

- Input parameters
  - Algorithm parameters
  - Problem size
  - etc.

- Initial states and intermediate states of the system for executing
  - Cache configuration, replacement policies
  - Pipelines
  - Speculations
  - etc.

- Interferences from the environment
  - Scheduling
  - Interrupts
  - etc.
How to Derive the Worst-Case Execution Time (WCET)

- Most of industry’s best practice
  - Measure it: determine WCET directly by running or simulating a set of inputs.
    - There is no guarantee to give an upper bound of the WCET.
    - The derived WCET could be too optimistic.
  - Exhaustive execution: by considering the set of all the possible inputs
    - In general, not possible
    - The inputs have to cover all the possible initial states and intermediate states of the system, which is also usually not possible.

- Compute it
  - In general, not possible neither, as computing (tight) WCET for a program is *uncomputable* by Turing machines.
  - Based on some structures, it is possible and the derived solution is a safe upper bound of the WCET.
Why is It Uncomputable?

**Halting Problem**

Given the description of a Turing machine $m$ and its input $x$, the problem is to answer whether the machine halts on $x$.

**Theorem**

Halting Problem is undecidable (uncomputable). In other words, one cannot use an algorithm to decide whether another algorithm $m$ halts on a specific input.

**WCET is undecidable**

It is even undecidable if it terminates at all. Deriving the WCET is of course of undecidable.

Please refer to the textbook of Computational Complexity by Prof. Papadimitriou.
Our objectives:

- **Upper bound** of execution time as tightly as possible.
- All control-flow paths, by considering all possible inputs.
- All paths through the architecture, resulting from the potential initial and assumed intermediate architectural states.
Timing Analysis

By considering systems, in general, with

- finite architectural configurations, finite input domains, and bounded loops and recursion,

WCET is computable.
Timing Analysis

By considering systems, in general, with

- finite architectural configurations, finite input domains, and bounded loops and recursion,

WCET is computable.

But....... the search space is too large to explore it exhaustively!
Why is It Hard for Analyzing WCET?

Execution time $e(i)$ of machine instruction $i$

- In the good old time:
  $e(i)$ is a constant $c$, which could be found in the data sheet
- Nowadays, especially for high-performance processors:
  $e(i)$ also depends on the (architectural) execution state $s$.

$$\min\{e(i, s) | s \in S\} \leq e(i) \leq \max\{e(i, s) | s \in S\},$$

where $S$ is the set of all states.

- Using $\max\{e(i, s) | s \in S\}$ is safe for WCET, but might be not tight since some states in $S$ might not be possible reached by some inputs.

- Execution history, resulting in a smaller set of reachable execution states, has to be enforced to improve the tightness of the analysis.
Variability of Execution Times

\[ x = a + b; \]

\[
\begin{align*}
\text{LOAD} & \quad r2, _a \\
\text{LOAD} & \quad r1, _b \\
\text{ADD} & \quad r3, r2, r1
\end{align*}
\]

PPC 755

![Bar chart showing execution time (clock cycles) for best and worst case scenarios.](image)

Wilhelm et al.
Timing Accidents and Penalties

- **Timing Accident**: cause for an increase of the execution time of an instruction
- **Timing Penalty**: the associated increase
- **Types of timing accidents**
  - Cache misses
  - Pipeline stalls
  - Branch mispredictions
  - Bus collisions
  - Memory refresh of DRAM
  - TLB miss
Overall Approach: Modularization

- Architecture Analysis:
  - Use Abstract Interpretation.
  - Exclude as many Timing Accidents as possible during analysis.
    - Certain timing accidents will never happen, e.g., at a certain program point, instruction fetch will never cause a cache miss.
    - The more accidents excluded, the lower (better) the upper bound.
  - Determine WCET for basic blocks, based on context information.

- Worst-Case Path Determination:
  - Map control flow graph to an Integer Linear Program (ILP).
  - Determine upper bound and associated path.

High-Level Objectives: Upper Bound of WCET

- It must be safe, i.e., not underestimate.
- It should be tight, i.e., not far away from real WCET.
- The analysis effort must be tolerable.
Overall Structure

Executable Binary Program

Control-Flow-Graph (CFG) Reconstruction

Loop Analysis and Unfolding

Loop Bounds

Static Analysis
- Value Analyzer
- Cache/Pipeline Analyzer

Path Analysis
- ILP-Generator
- ILP-Solver
- Evaluation

Micro-architecture Analysis
- WCET Visualization and Analysis

Worst-Case Path Analysis

Prof. Dr. Jian-Jia Chen (LS 12, TU Dortmund)
Outline

Introduction

Program Path Analysis

Static Analysis
  Value Analysis
  Cache Analysis
what_is_this {  
    read (a,b);  
    done = FALSE;  
    repeat {  
        if (a>b) {  
            a = a-b;  
        } elseif (b>a) {  
            b = b-a;  
        } else done = TRUE;  
    } until done;  
    write (a);  
}
Definition: A basic block is a sequence of instructions where the control flow enters at the beginning and exits at the end, in which it is highly amenable to analysis.

\[
a[0] := b[0] + c[0] \\
d := a[0] \times a[1] \\
e := d / a[2] \\
\text{if } e < 10 \text{ goto } L
\]
Basic Blocks

Definition: A basic block is a sequence of instructions where the control flow enters at the beginning and exits at the end, in which it is highly amenable to analysis.

Determining the basic blocks

- Beginning:
  - the first instruction
  - targets of un/conditional jumps
  - instructions that follow un/conditional jumps

- Ending:
  - the basic block consists of the block beginning and runs until the next block beginning (exclusive) or until the program ends

```
a[0] := b[0] + c[0]
d := a[0] * a[1]
e := d / a[2]
if e < 10 goto L
```
Program Path Analysis

- Problem: Which sequence of instructions is executed in the worst case (i.e., the longest execution time)?
- Input:
  - Timing information for each basic block, derived from static analysis (value/cache/pipeline analysis)
  - Loop bounds by specification
  - CFG derived from the executable binary program
- Basic Concept:
  - Transform structure of CFG into a set of (integer) linear equations
  - Solution of the Integer Linear Program (ILP) yields bound on the WCET.
Program Path Analysis: Formal Definition

Input

A CFG with $N$ basic blocks, in which each basic block $B_i$ has a worst-case execution time $c_i$, given by static analysis.

Output

Suppose that each block $B_i$ is executed exactly $x_i$ times. What is the worst-case execution time

$$WCET = \sum_{i=1}^{N} c_i \cdot x_i,$$

such that the values of $x_i$s satisfy the structural constraints in the CFG?

Note that additional constraints provided by the programmer (bounds for loop counters, etc.) can also be included.
Example for CFG Constraints

Flow equations: ($x_i$ is a variable)

\[
\begin{align*}
    d_1 &= d_2 = x_1 \\
    d_3 + d_9 &= d_2 + d_8 = x_2 \\
    d_4 + d_5 &= d_3 = x_3 \\
    d_6 + d_7 &= d_8 = x_4 \\
    d_4 &= d_6 = x_5 \\
    d_5 &= d_7 = x_6
\end{align*}
\]
The loop is executed for at most 20 times when \( k \) is initialized with a non-negative number:

\[ x_3 \leq 20x_1. \]

The basic block for \( j = 0; ok = true; \) is executed for at most one time:

\[ x_6 \leq x_1. \]
WCET: ILP Formulation

maximize \[ \sum_{i=1}^{N} c_i \cdot x_i \]

such that \[ d_1 = 1 \]

\[ \sum_{j \in \text{in}(B_i)} d_j = \sum_{k \in \text{out}(B_j)} d_k = x_i, \quad \forall i = 1, \ldots, N \]

additional linear constraints
Outline

Introduction

Program Path Analysis

Static Analysis
  Value Analysis
  Cache Analysis
Overall Structure

Executable Binary Program

Control-Flow-Graph (CFG) Reconstruction

Loop Analysis and Unfolding

Loop Bounds

Static Analysis

Value Analyzer

Cache/Pipeline Analyzer

Micro-Architecture Abstraction

Timing Information

Path Analysis

ILP-Generator

ILP-Solver

Evaluation

WCET Visualization and Analysis

Micro-architecture Analysis

Worst-Case Path Analysis
Outline

Introduction

Program Path Analysis

Static Analysis
- Value Analysis
- Cache Analysis
Value Analysis: Motivation and Method

- **Motivation**
  - Provide access information to data-cache/pipeline analysis
  - Detect infeasible paths
  - Derive loop bounds

- **Method**
  - Calculate intervals at all program points
  - By considering addresses, register contents, local and global variables.

**Abstract Interpretation**

Perform the program’s computation using value descriptions or abstract values in place of the concrete values.
Outline

Introduction

Program Path Analysis

Static Analysis
  Value Analysis
  Cache Analysis
Caches: Fast Memory to Deal with the Memory Wall

- How they work:
  - dynamically
  - managed by replacement policy

![Diagram of CPU, Cache, and Main Memory]

- Why they work: *principle of locality*
  - spatial
  - temporal
Caches: Fast Memory to Deal with the Memory Wall

- How they work:
  - dynamically
  - managed by replacement policy

```
CPU -> Cache -> Main Memory
```

- Capacity: 32 KB
- Latency: 3 cycles
- Main Memory: 2 MB
- Latency: 100 cycles

- Why they work: principle of locality
  - spatial
  - temporal
Caches: Fast Memory to Deal with the Memory Wall

• How they work:
  • dynamically
  • managed by replacement policy

![Diagram of CPU, Cache, and Main Memory with "hit" and latency information]

• Why they work: principle of locality
  • spatial
  • temporal

Capacity: 32 KB
Latency: 3 cycles
Main Memory: 2 MB
Latency: 100 cycles
Caches: Fast Memory to Deal with the Memory Wall

- How they work:
  - dynamically
  - managed by replacement policy

- Why they work: principle of locality
  - spatial
  - temporal

![Diagram of CPU, Cache, and Main Memory with "miss" and latency information]
Caches: Fast Memory to Deal with the Memory Wall

- How they work:
  - dynamically
  - managed by replacement policy

```
CPU                      Cache                      Main Memory
```

```
Capacity: 32 KB
Latency: 3 cycles
```

```
Capacity: 2 MB
Latency: 100 cycles
```

- Why they work: principle of locality
  - spatial
  - temporal
Caches: Fast Memory to Deal with the Memory Wall

- **How they work:**
  - dynamically
  - managed by replacement policy

![Diagram of CPU, Cache, and Main Memory]

- **Capacity:** 32 KB
- **Latency:** 3 cycles
- **Main Memory:** 2 MB
- **Latency:** 100 cycles

- **Why they work:** *principle of locality*
  - spatial
  - temporal
Caches: Fast Memory to Deal with the Memory Wall

- **How they work:**
  - dynamically
  - managed by replacement policy

- **Why they work:** *principle of locality*
  - spatial
  - temporal

**Diagram:**

- CPU
- Cache
- Main Memory

**Specifications:**

- **Capacity:**
  - Cache: 32 KB
  - Main Memory: 2 MB

- **Latency:**
  - Cache: 3 cycles
  - Main Memory: 100 cycles
Caches: Fast Memory to Deal with the Memory Wall

- How they work:
  - dynamically
  - managed by replacement policy

- Why they work: principle of locality
  - spatial
  - temporal

```
CPU ➔ Cache ➔ Main Memory
```

```
Capacity: 32 KB
Latency: 3 cycles
```

```
2 MB
100 cycles
```

"hit" [ac]
Caches: Fast Memory to Deal with the Memory Wall

- How they work:
  - dynamically
  - managed by replacement policy

- Why they work: *principle of locality*
  - spatial
  - temporal

**Diagram:**

- CPU
- Cache
- Main Memory

**Capacity:**
- Cache: 32 KB
- Main Memory: 2 MB

**Latency:**
- Cache: 3 cycles
- Main Memory: 100 cycles

“hit” [ac]
Fully Associative Caches

Address:

Tag

Block
offset

Tag
Data Block

Tag
Data Block

... 

Tag
Data Block

MUX

Data

=?

Yes:
Hit!

No:
Miss!

$log_2(8 \times b)$
Set-Associative Caches

Special cases:
- direct-mapped cache: only one line per cache set
- fully-associative cache: only one cache set
Replacement Policies

- Least-Recently-Used (LRU) used in **Intel Pentium I** and **MIPS 24K/34K**
- First-In First-Out (FIFO or Round-Robin) used in **Motorola PowerPC 56x**, **Intel XScale**, **ARM9**, **ARM11**
- Pseudo-LRU (PLRU) used in **Intel Pentium II-IV** and **PowerPC 75x**
- Most Recently Used (MRU) as described in literature

Each cache set is treated independently:

→ Set-associative caches are compositions of fully-associative caches.
Cache Analysis for LRU

Two types of cache analyses:

1. **Local guarantees: classification of individual accesses**
   - Must-Analysis $\rightarrow$ Underapproximates cache contents
   - May-Analysis $\rightarrow$ Overapproximates cache contents

2. **Global guarantees: bounds on cache hits/misses**
Challenges for Cache Analysis

Always a cache hit/always a miss?

1. Initial cache contents unknown.
2. Different paths lead to these points.
3. Cannot resolve address of $z$. 

Prof. Dr. Jian-Jia Chen (LS 12, TU Dortmund)
Challenges for Cache Analysis

Always a cache hit/always a miss?

1. Initial cache contents unknown.
2. Different paths lead to these points.
3. Cannot resolve address of $z$. 

read $z$

read $y$

read $x$

write $z$
Using Abstract Interpretation

Collecting Semantics =
set of states at each program point that
any execution may encounter there

Two approximations:

Collecting Semantics  uncomputable
⊆ Cache Semantics  computable
⊆ γ(Abstract Cache Sem.) efficiently com-
putable
Using Abstract Interpretation

Collecting Semantics =
set of states at each program point that
any execution may encounter there

Two approximations:

Collecting Semantics uncomputable
\subseteq Cache Semantics computable
\subseteq \gamma (Abstract Cache Sem.) efficiently com-
putable
Using Abstract Interpretation

\[
\text{Collecting Semantics} = \text{set of states at each program point that any execution may encounter there}
\]

Two approximations:

\[
\text{Collecting Semantics} \subseteq \text{Cache Semantics} \subseteq \gamma(\text{Abstract Cache Sem.})
\]

uncomputable

computable

efficiently computable
Using Abstract Interpretation

Collecting Semantics = set of states at each program point that any execution may encounter there

Two approximations:

Collecting Semantics \subseteq \text{Cache Semantics} \subseteq \gamma(\text{Abstract Cache Sem.})

uncomputable
computable
efficiently computable
Using Abstract Interpretation

Collecting Semantics = set of states at each program point that any execution may encounter there

Two approximations:

Collecting Semantics uncomputable

⊆ Cache Semantics computable

⊆ γ(Abstract Cache Sem.) efficiently computable
Least-Recently-Used (LRU): Concrete Behavior

“Cache Miss”:

LRU has notion of age

“Cache Hit”:
LRU: Must-Analysis: Abstract Domain

- Used to predict *cache hits*.
- Maintains *upper bounds on ages* of memory blocks.
- Upper bound $\leq$ associativity $\rightarrow$ memory block definitely cached.

**Example**

Describes the set of all concrete cache states in which $x$, $s$, and $t$ occur,
- $x$ with an age of 0,
- $s$ and $t$ with an age not older than 2.

$$
\gamma([\{x\}, \{\}, \{s, t\}, \{\}]) = \\
\{[x, s, t, a], [x, t, s, a], [x, s, t, b], \ldots\}
$$
Sound Update – Local Consistency

\[ (\text{must}) \xrightarrow{\text{Abstract Update}} (\text{must'}) \]

\[ \gamma \]

Concrete cache states

Lifted Concrete Update

Concrete cache states
LRU: Must-Analysis: Update

"Potential Cache Miss":

"Definite Cache Hit":

Why does not $t$ age in the second case under Must-Cache analysis?
LRU: Must-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

```
{a}  {c}
{d}  {}
{c,f}
{e}  {}
{d}
{a}  
{a,c} 
```

"Intersection + Maximal Age"
**LRU: Must-Analysis: Join**

Need to combine information where control-flow merges.

Join should be conservative:

- \( \gamma(A) \subseteq \gamma(A \sqcup B) \)
- \( \gamma(B) \subseteq \gamma(A \sqcup B) \)

```
\{a\}  \{c\}  \{\}  \\
\{\}  \{e\}  \{\}  \\
\{c,f\}  \{a\}  \{a,c\}  \\
\{d\}  \{d\}  \{d\}  
```

“Intersection + Maximal Age”
LRU: Must-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

```
\begin{array}{c}
\{a\} \\
\{\} \\
\{c,f\} \\
\{d\}
\end{array} \sqcup
\begin{array}{c}
\{c\} \\
\{e\} \\
\{a\} \\
\{d\}
\end{array} =
\begin{array}{c}
\{} \\
\{} \\
\{a,c\} \\
\{d\}
\end{array}
```

"Intersection + Maximal Age"
LRU: Must-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

```
{{a}}
{{}}
{{c,f}}
{{d}}

{{}}
{{e}}
{{a}}
{{d}}

{{}}
{{}}
{{a,c}}
{{d}}
```

“Intersection + Maximal Age”
LRU: Must-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:
- \( \gamma(A) \subseteq \gamma(A \sqcup B) \)
- \( \gamma(B) \subseteq \gamma(A \sqcup B) \)

\[
\begin{array}{ccc}
\{a\} & \{c\} & \{\} \\
\{\} & \{e\} & \{\} \\
\{c,f\} & \{a\} & \{a,c\} \\
\{d\} & \{d\} & \{d\} \\
\end{array}
\]

“Intersection + Maximal Age”
LRU: Must-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- \( \gamma(A) \subseteq \gamma(A \sqcup B) \)
- \( \gamma(B) \subseteq \gamma(A \sqcup B) \)

\[
\begin{array}{c}
\{a\} \\
\{} \\
\{c,f\} \\
\{d\}
\end{array}
\quad \sqcup \quad
\begin{array}{c}
\{c\} \\
\{e\} \\
\{a\} \\
\{d\}
\end{array}
\quad \begin{array}{c}
\{} \\
\{} \\
\{a,c\} \\
\{d\}
\end{array}
\]

“Intersection + Maximal Age”
LRU: Must-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

```
\{a\}       \{c\}       \{\}
\{\}        \{e\}        \{
\{c,f\}     \{a\}        \{\}
\{d\}       \{d\}        \{a,c\}
```

“Intersection + Maximal Age”

How many memory blocks can be in the must-cache?
LRU: May-Analysis: Abstract Domain

- Used to predict *cache misses*.
- Maintains *lower bounds on ages* of memory blocks.
- Lower bound $\geq$ associativity
  $\rightarrow$ memory block definitely *not* cached.

Abstract state:

<table>
<thead>
<tr>
<th>Age 0</th>
<th></th>
<th>Age 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>{x,y}</td>
<td></td>
<td>{x,y}</td>
</tr>
<tr>
<td>{}</td>
<td></td>
<td>{}</td>
</tr>
<tr>
<td>{s,t}</td>
<td></td>
<td>{u}</td>
</tr>
</tbody>
</table>

and its interpretation:

Describes a set of all concrete cache states, where no memory blocks except $x$, $y$, $s$, $t$, and $u$ occur,
- $x$ and $y$ with an age of at least 0,
- $s$ and $t$ with an age of at least 2,
- $u$ with an age of at least 3.

$\gamma([\{x, y\}, \{\}, \{s, t\}, \{u\}]) =$

\{ $[x, y, s, t], [y, x, s, t], [x, y, s, u], \ldots$ \}
LRU: May-Analysis: Update

“Definite Cache Miss”:

“Potential Cache Hit”:

Why does $t$ age in the second case?
LRU: May-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

"Union + Minimal Age"
LRU: May-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

\[
\begin{array}{ccc}
\{a\} & \{c\} & \{a,c\} \\
\{\} & \{e\} & \{e\} \\
\{c,f\} & \{a\} & \{f\} \\
\{d\} & \{d\} & \{d\}
\end{array}
\]

“Union + Minimal Age”
LRU: May-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

```
{a}   {c}   {a,c}
{c,f} {e}   {e}
{d}   {a}   {d}
{d}   {f}   {f}
{e}   {d}   {d}
```

“Union + Minimal Age”
LRU: May-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

```
\begin{array}{c|c|c}
\{a\} & \{c\} & \{a,c\} \\
\{\} & \{e\} & \{e\} \\
\{c,f\} & \{a\} & \{f\} \\
\{d\} & \{d\} & \{d\}
\end{array}
```

“Union + Minimal Age”
LRU: May-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

```
\begin{array}{c}
\{a\} \\
\{\} \\
\{c,f\} \\
\{d\}
\end{array}
\sqcup
\begin{array}{c}
\{c\} \\
\{e\} \\
\{a\} \\
\{d\}
\end{array}
= 
\begin{array}{c}
\{a,c\} \\
\{e\} \\
\{f\} \\
\{d\}
\end{array}
```

“Union + Minimal Age”
LRU: May-Analysis: Join

Need to combine information where control-flow merges.

Join should be conservative:

- $\gamma(A) \subseteq \gamma(A \sqcup B)$
- $\gamma(B) \subseteq \gamma(A \sqcup B)$

```
\{a\} \sqcup \{c, f\} \sqcup \{d\} = \{a, c\} \sqcup \{e\} \sqcup \{f\} \sqcup \{d\}
```

“Union + Minimal Age”
Summary of WCET Analysis

- Value analysis
- Cache analysis
  - using statically computed effective addresses and loop bounds
- Pipeline analysis
  - Model CPU as a big state machine.
  - There have been some new results published from Prof. Jan Reineke’s group from Uni des Saarlandes.
aiT-Tool